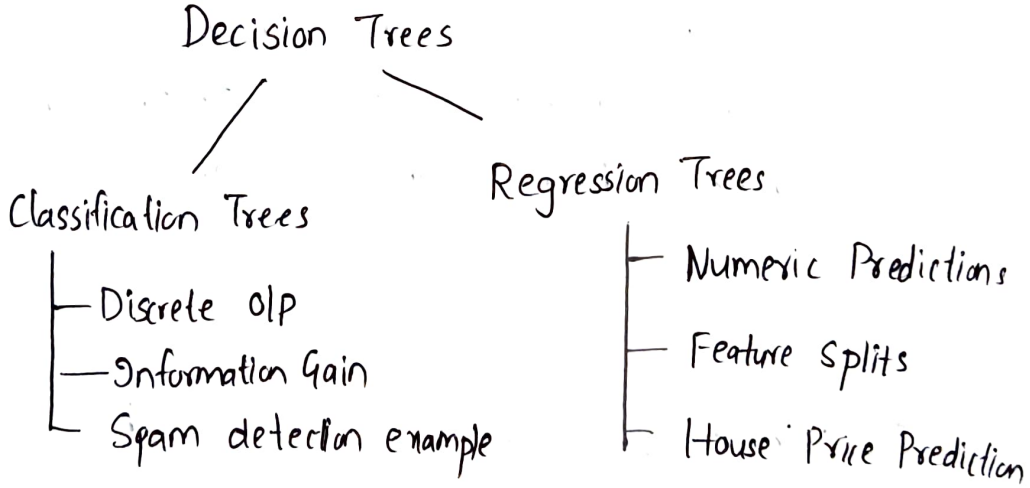
UNIT - II

Decision Tree

Decision trees are a simple machine learning tool used for classification and regression tasks. They break complex decisions into smaller steps, making them easy to understand and implement.

A decision tree, which has a hierarchical structure made up of root, branches, internal and leaf nodes is a supervised learning approach used for classification and regression applications.

Types of decision trees



Classification Trees:

→ These trees are used for categorical target variables where o/p is discrete such as "Yes/No" or "Spam/Not spam".

→ The tree is built by splitting the dataset based on feature values that maximize information gain.

2) Regression Trees:

→ These are used when the target variable is continuous such as predicting house prices or stock market values.

→ It predicts a numeric value based on feature splits.

Advantages:

→ Easy to understand
→ Feature selection is inbuilt
→ Can handle missing values
→ Requires minimal data preprocessing
→ Handles both numerical and categorical data

Limitations:

1) Overfitting: A decision tree with too many splits may fit the training data perfectly but fail to generalize well to unseen data.

2) Unstable: Even small changes in the dataset can result in significantly different tree structure, making decision trees sensitive to noise.

3) Biased towards dominant classes

4) Computationally expensive.

Applications :

1) Medical Diagnosis
2) Customer Segmentation
3) Fraud Detection
4) Recommendation Systems
5) Credit Scoring

Types of decision tree algorithms

1) ID3: (Iterative Dichotomiser 3)

It is a classic decision tree algorithm commonly used for classification tasks. It calculates entropy and information gain for each feature and selects the feature with highest information gain for splitting.

ID3 recursively splits the dataset using the feature with the highest information gain until all examples in a node belong to the same class or no features remain to split. After the tree is constructed it prune branches that don't significantly improve accuracy to reduce overfitting. But it tends to overfit the training data and cannot directly handle continuous attributes. These issues are addressed by other algorithms like C4.5 and CART.

2) C4.5

It uses a modified version of information gain called the gain ratio to reduce the bias towards features with many values

Advantages and Limitations of ID3

Advantages:

1) Handles categorical data
2) Interpretability
3) Computationally inexpensive.

Limitations:

1) Overfitting
2) Sensitive to noise
3) Binary trees only (ID3 constructs binary trees, limiting its ability to represent more complex relationships present in data directly)

NOTE:
The goal of ID3 is to maximize the homogeneity (reduce impurity) of the resulting subsets.

Why the name ID3

ID3 stands for Iterative Dichotomiser 3.

Iterative: The algorithm repeatedly (iteratively) goes through the process of selecting the best attribute to split the data.

Dichotomiser: It dichotomizes or divides, the data into two or more subsets at each step based on the values of the chosen attribute.

3: It is the third in a series of algorithms developed by the computer scientist Ross Quinlan following earlier work.

Entropy: Entropy measures the impurity or uncertainty in a dataset. It is calculated using the formula

$$\text{Entropy} = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}$$

Information Gain: It measures the reduction in entropy achieved by splitting the dataset based on a feature. It is called as

$$\text{Information Gain}(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \cdot H(D_v)$$

$H(D_v)$ → entropy of subset of instances where feature A has value v

$|D|$ → Total number of instances in dataset D

The goal of ID3 is to maximize the homogeneity (reduce impurity) of the resulting subsets.

2) C4.5

It uses a modified version of information gain called the gain ratio to reduce the bias towards features with many values.

Instead of information gain, C4.5 uses gain ratio as the splitting criterion, which normalizes the information gain by the intrinsic value of the feature

It is calculated as

$$\text{Gain Ratio}(A) = \frac{\text{Information Gain}(A)}{\text{Split Information}(A)}$$

A – feature

Split Information (A): It is a measure of intrinsic information of feature A also known as its split entropy or intrinsic value.

$$\text{Split Information}(A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log_2 \frac{|D_i|}{|D|}$$

n – no. of distinct values for feature A

$|D|$ – Tot. no. of instances in dataset

$|D_i|$ → no. of instances in dataset where feature A takes on i^{th} value.

Gini index (D) = $1 - \sum_{i=1}^{c} P_i^2$

C → no. of classes

P_i → proportion of samples in class 'i' in dataset D.

For regression CART builds regression trees by minimizing the variance of target variable within each subset. The split that reduces the variance the most is chosen.

To reduce overfitting CART uses cost-complexity pruning after tree construction. It involves minimizing a cost function that combines the impurity and tree complexity by adding a complexity parameter to the impurity measure.

4) CHAID (Chi-Square Automatic Interaction Detection)

Used to determine the best splits especially for categorical variables. It recursively divides the data into smaller subsets until each subset contains only data points of the same class within a specified range of values.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

O_i → observed frequency

E_i → expected frequency in each category.

Limitations of C4.5

→ It can be prone to overfitting especially in noisy datasets

→ Performance may degrade when dealing with datasets that have many features.

3) CART (Classification and Regression Trees)

It can handle both classification and regression tasks. It uses the Gini index as the splitting criterion for classification tasks and reduction in variance for regression tasks.

Gini index:

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly Classified when it is randomly Chosen. It considers a binary split for each attribute. The attribute with smallest Gini index is chosen to split the data.

Gini index ranges from 0 to 0.5 where

0 → dataset is perfectly pure, meaning all samples belong to the same class.

0.5 → indicates maximum impurity, meaning the dataset is evenly split among all classes

Decision Tree - ID3 Algorithm

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Need to calculate information gain.
The attribute which is having highest information gain will be taken as root node.

Attribute : Outlook

The possible values are Sunny, Overcast, Rain

Values (Outlook) = Sunny, overcast, Rain

Before calculating Information gain, calculate entropy of whole attribute and individual entropy.

Entropy of whole dataset

$S = [9+, 5-]$ — 9 +ve examples
5 -ve examples.

$$\text{entropy}(S) = \underset{0.12}{\frac{-9}{14}\log_2\frac{9}{14}} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

$$\text{Entropy}(S) = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}$$

Need to calculate entropy of Sunny, overcast, Rain

Entropy of Sunny

D_1	Sunny	NO
D_2	Sunny	NO
D_8	Sunny	NO
D_9	Sunny	yes
D_{11}	Sunny	yes

$S_{sunny} \leftarrow [2+, 3-]$

$$\text{entropy}(S_{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.971$$

$S_{Overcast}$

D3	Yes
D7	Yes
D12	Yes
D13	Yes

$S_{Overcast} \leftarrow [4+, 0-]$

$Entropy(S_{overcast}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}$

$= 0$

S_{Rain}

D4	Yes
D5	Yes
D6	NO
D10	Yes
D14	NO

$S_{Rain} \leftarrow [3+, 2-]$

$Entropy(S_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.971$

$Gain(S, Outlook) = Entropy(S) - \sum_{v \in (Sunny, Overcast, Rain)} \frac{|S_v|}{|S|} Entropy(S_v)$

$= Entropy(S) - \frac{3+3}{14} Entropy(S_{Sunny}) - \frac{4+0}{14} Entropy(S_{Overcast}) - \frac{3+2}{14} Entropy(S_{Rain})$

NOTE: 1) If no. of positive examples and negative examples are equal, then entropy is **1.0**

2) If we have either only positive examples or only negative examples, then entropy is **0**

Gain (S, Outlook) = $0.94 - \frac{5}{14}(0.971) - \frac{4}{14}(0)$ -

$\frac{5}{14}(0.971) = 0.2464$ ∴

Gain (S, Outlook) = 0.2464

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

S = [9+, 5-] Entropy(S) = $\frac{-9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

S = 0.94

Hot - Entropy

D_1	NO
D_2	NO
D_3	Yes
D_{13}	Yes

$S_{Hot} \leftarrow [2+, 2-]$

Entropy (S_{Hot}) = $\frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$

= 1.0

Mild

D_4	Yes
D_8	NO
D_{11}	Yes
D_{12}	Yes
D_{14}	NO
D_{10}	Yes

$S_{Mild} \leftarrow [4+, 2-]$

Entropy (S_{Mild}) = $\frac{-4}{6}\log_2\frac{4}{6} - \frac{2}{3}\log_2\frac{2}{6}$

= 0.9183

Cool

D5	Yes
D6	NO
D7	yes
D9	yes

$S_{cool} \leftarrow [3+, 1-]$

$Entropy(S_{cool}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 0.8113$

$Gain(S, Temp) = Entropy(S) - \sum_{v \in (Hot, cool, Mild)} \frac{|S_v|}{|S|} Entropy(S_v)$

$= Entropy(S) - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild}) - \frac{4}{14} Entropy(S_{cool})$

$= 0.94 - \frac{4}{14}(1.0) - \frac{6}{14}(0.9183) - \frac{4}{14}(0.8113)$

$= 0.0289$

Attribute : Humidity

Values (Humidity) = High, Normal

$S = [9+, 5-]$ $Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

S_{High}

D1	NO
D2	NO
D3	yes
D4	yes
D8	NO
D12	yes
D14	NO

$S_{High} \leftarrow [3+, 4-]$

$Entropy(S_{High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$

S_{Normal}

D_5	yes
D_6	No
D_7	yes
D_9	yes
D_{10}	yes
D_{11}	yes
D_{13}	yes

$S_{Normal} \leftarrow [6+, 1-]$

$Entropy(S_{Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$

$= 0.5916$

$Gain(S, Humidity) = Entropy(S) - \sum_{v \in (High, Normal)} \frac{|S_v|}{|S|} Entropy(S_v)$

$= Entropy(S) - \frac{7}{14} Entropy(S_{High}) - \frac{7}{14} Entropy(S_{Normal})$

$= 0.94 - \frac{7}{14}(0.9852) - \frac{7}{14}(0.5916$

$= 0.1516$

Attribute : Wind

Values (Wind) = strong, weak

$S = [9+, 5-]$ $Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

$S_{strong} \leftarrow [3+, 3-] = Entropy(S_{strong}) = 1.0$

$S_{weak} \leftarrow [6+, 2-] = Entropy(S_{weak}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$= 0.8113$

$Gain(S, Wind) = Entropy(S) - \sum_{v \in (strong, weak)} \frac{|S_v|}{|S|} Entropy(S_{weak})$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{strong}) - \frac{8}{14} Entropy(S_{weak})$$

$$= 0.94 - \frac{6}{14}(1.0) - \frac{8}{14}(0.8113) = 0.0478$$

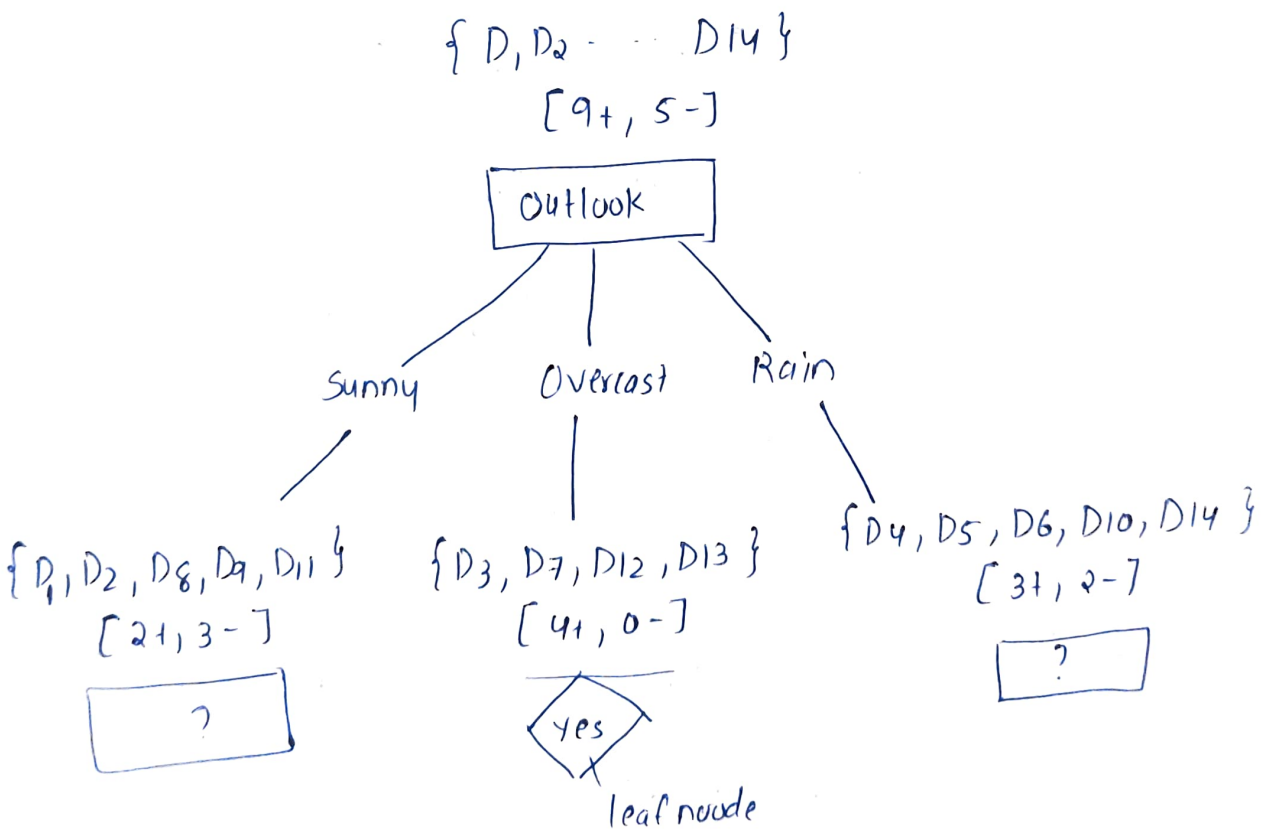
Gain (S, Outlook) = 0.2464

Gain (S, Temp) = 0.0289

Gain (S, Humidity) = 0.1516

Gain (S, Wind) = 0.0478

Outlook is having maximum gain. So consider this as root node.



We need to calculate Information gain again for Sunny and Rain as we didn't get any leaf node.

Sunny

Day	Temp	Humidity	Wind	Play Tennis
D_1	Hot	High	Weak	NO
D_2	Hot	High	Strong	NO
D_8	Mild	High	Weak	NO
D_9	Cool	Normal	Weak	Yes
D_{11}	Mild	Normal	Strong	Yes

Attribute : Temp

values (Temp) = Hot, Mild, Cool

S_{sunny} = [2+, 3-] Entropy (S_{sunny}) = $\frac{-2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5}$

= 0.97

$S_{Hot} \leftarrow$ [0+, 2-] Entropy (S_{Hot}) = 0.0

$S_{Mild} \leftarrow$ [1+, 1-] ; Entropy (S_{Mild}) = 1.0

$S_{Cool} \leftarrow$ [1+, 0-] Entropy (S_{Cool}) = 0.0

$$Gain(S_{sunny}, Temp) = Entropy(S) - \sum_{v \in (Hot, Mild, Cool)} \frac{|S_v|}{|S|} Entropy(S_v)$$

= Entropy (S) - $\frac{2}{5}$ Entropy (S_{Hot}) - $\frac{2}{5}$ Entropy (S_{Mild}) - $\frac{1}{5}$ Entropy (S_{Cool})

= 0.97 - $\frac{2}{5}$ (0) - $\frac{2}{5}$ (1) - $\frac{1}{5}$ (0)

= 0.570

Attribute : Humidity

Values (Humidity) = High, Normal

$S_{sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{High} \leftarrow [0+, 3-]$ $\quad Entropy(S_{High}) = 0.0$

$S_{Normal} \leftarrow [2+, 0-]$ $\quad Entropy(S_{Normal}) = 0.0$

$$Gain(S_{sunny}, Humidity) = Entropy(S) - \sum_{v \in (High, Normal)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{sunny}, Humidity) = Entropy(S) - \frac{3}{5} Entropy(S_{High}) - \frac{2}{5} Entropy(S_{Normal})$$

$$= 0.97 - \frac{3}{5}(0.0) - \frac{2}{5}(0.0) = 0.97$$

Attribute : Wind

Values (Wind) = Strong, Weak

$S_{sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$S_{strong} \leftarrow [1+, 1-]$ $\quad Entropy(S_{strong}) = 1.0$

$S_{weak} \leftarrow [1+, 2-]$ $\quad Entropy(S_{weak}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9183$

$$Gain(S_{sunny}, Wind) = Entropy(S) - \frac{2}{5} Entropy(S_{strong}) - \frac{3}{5} Entropy(S_{weak})$$

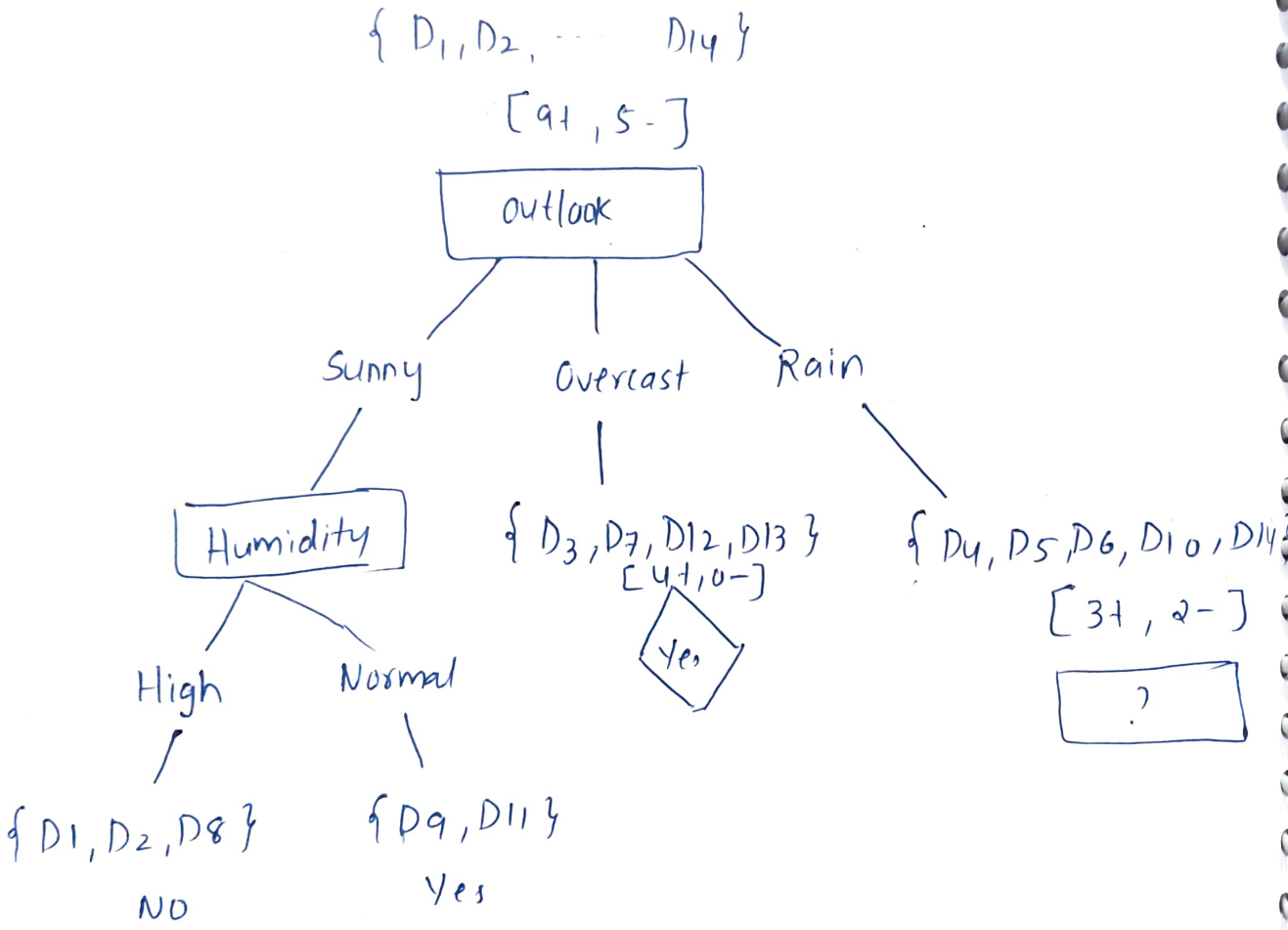
$$= 0.97 - \frac{2}{5}(1.0) - \frac{3}{5}(0.918) = 0.0192$$

Gain (S_{sunny} , Temp) : 0.570

Gain (S_{sunny} , Humidity) = 0.97

Gain (S_{sunny} , Wind) = 0.0192

Gain (S_{sunny} , Humidity) = 0.97 → node



Rain

Day	Temp	Humidity	Wind	Play Tennis
D4	Mild	High	Weak	Yes
D5	Cool	Normal	~~Strong~~ Weak	Yes
D6	Cool	Normal	~~Weak~~ Strong	NO
D10	Mild	Normal	~~Str~~ Weak	Yes
D14	Mild	High	Strong	NO

Attribute (Temp)

Values (Temp) = Hot, Mild, Cool

$S_{Rain} = [3+, 2-]$ $Entropy(S_{sunny}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

$S_{Hot} \leftarrow [0+, 0-]$ $Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [2+, 1-]$ $Entropy(S_{Mild}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9183$

$S_{Cool} \leftarrow [1+, 1-]$ $Entropy(S_{Cool}) = 1.0$

$$Gain(S_{Rain}, Temp) = Entropy(S) - \sum_{v \in (Hot, Mild, Cool)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp) = Entropy(S) - \frac{0}{5} Entropy(S_{Hot}) - \frac{3}{5} Entropy(S_{Mild}) - \frac{2}{5} Entropy(S_{Cool})$$

$$= 0.97 - \frac{0}{5}(0.0) - \frac{3}{5}(0.918) - \frac{2}{5}(1.0) = 0.0192$$

Attribute : Humidity

Values (Humidity) = High, Normal

$S_{Rain} = [3+, 2-]$ Entropy $(S_{sunny}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= 0.97$

$S_{High} \leftarrow [1+, 1-]$ Entropy $(S_{High}) = 1.0$

$S_{Normal} \leftarrow [2+, 1-]$ Entropy $(S_{normal}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

$= 0.9183$

$$Gain(S_{Rain}, Humidity) = Entropy(S) - \sum_{v \in (High, Normal)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - \frac{2}{5} Entropy(S_{High}) - \frac{3}{5} Entropy(S_{Normal})$$

$$= 0.97 - \frac{2}{5}(1.0) - \frac{3}{5}(0.918) = 0.0192$$

Attribute : Wind

Values (wind = strong, weak

$S_{Rain} = [3+, 2-]$ Entropy $(S_{sunny}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{3}{5}$

$= 0.97$

$S_{strong} \leftarrow [0+, 2-]$ Entropy $(S_{strong}) = 0.0$

$S_{weak} \leftarrow [3+, 0-]$ Entropy $(S_{weak}) = 0.0$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \sum_{v \in (strong, weak)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5} Entropy(S_{strong}) - \frac{3}{5} Entropy(S_{weak})$

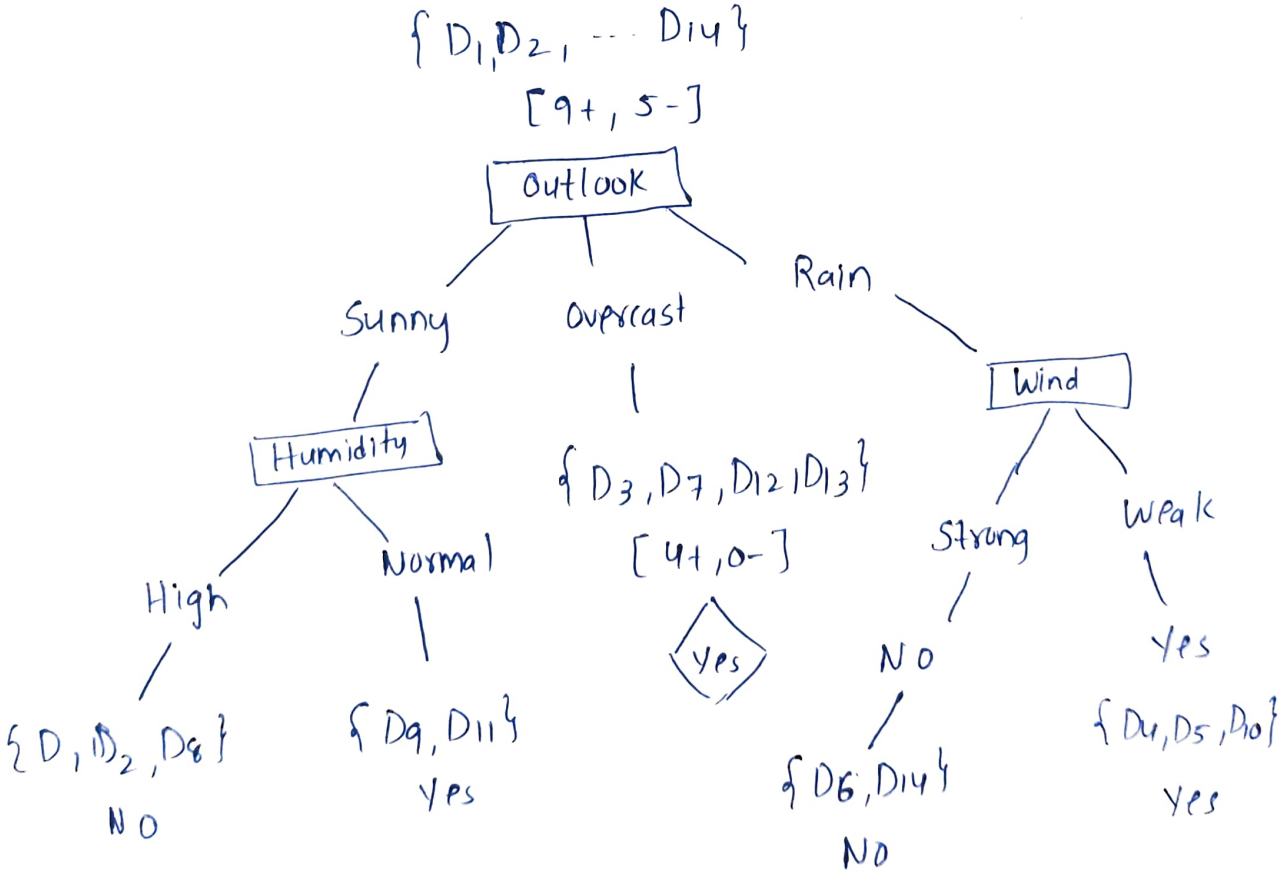
$= 0.97 - \frac{2}{5}(0.0) - \frac{3}{5}(0.0) = 0.97$

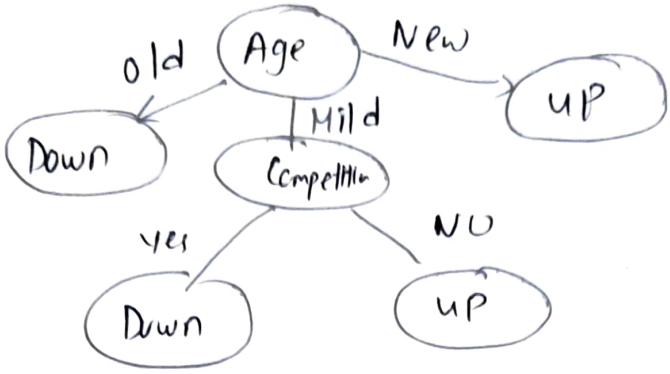
$Gain(S_{Rain}, Temp) = 0.0192$

$Gain(S_{Rain}, Humidity) = 0.0192$

$Gain(S_{Rain}, Wind) = 0.97$

0.97 is more, so wind is considered as node





Attribute : Age

S → E = 1

S_{old} E = 0

S_{Mid} E = 1

S_{New} E = 0

Gain = 0.6

Attribute : Competition

EXAMPLE PROBLEMS

Decision Tree Algorithm - ID3

1) What is the entropy of this collection of training examples w.r.t the target function classification ?

2) What is the information gain of a_1 and a_2 relative to these training examples ?

3) Draw decision tree for the given dataset.

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	−	T	F
4	+	F	F
5	−	F	T
6	−	F	T

Attribute : a_1

values (a_1) = T, F

$S = [3+, 3-]$ Entropy $(S) = 1.0$

$S_T \leftarrow [2+, 1-]$ Entropy $(S_T) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{4}$

$= 0.9183$

$S_F \leftarrow [1+, 2-]$ Entropy $(S_F) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9183$

$$\text{Gain}(S, a_1) = \text{Entropy}(S) - \sum_{v \in (T,F)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$$= 1.0 - \frac{3}{6} \times 0.9183 - \frac{3}{6} \times 0.9183$$

$$= 0.0817$$

Attribute : a_2

Values (a_2) = T, F

$S = [3+, 3-]$ Entropy $(S) = 1.0$

$S_T \leftarrow [2+, 2-]$ Entropy $(S_T) = 1.0$

$S_F \leftarrow [1+, 1-]$ Entropy $(S_F) = 1.0$

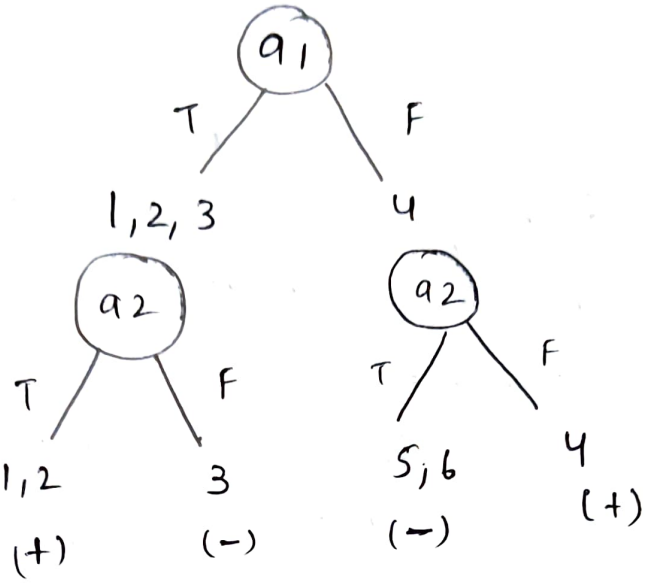
$$\text{Gain}(S, a_2) = \text{Entropy}(S) - \frac{4}{6} \text{Entropy}(S_T) - \frac{2}{3} \text{Entropy}(S_F)$$

$$= 1.0 - \frac{4}{6} \times 1.0 - \frac{2}{6} \times 1.0$$

$$= 0.0$$

Gain $(S, a_1) = 0.0817$ — Maximum Gain

Gain $(S, a_2) = 0.0$



As we have only one attribute, directly we are taking that attribute without calculating entropy, information gain.

Regression:

It is a supervised machine learning and statistical technique used to predict a continuous numerical value based on one or more i/p variables.

It finds the relationship blw variables so we can estimate or predict a number.

Ex: If you want to predict

- → House price based on size, location, no. of rooms
- → Student marks based on hours studied
- → Temperature based on time of day

These are regression problems because the o/p is a number, not a class.

(OR)

It is a process of determining a relationship blw one or more independent variables and one dependent or output variable.

Based on type of functions used to represent the relationship b/w dependent or o/p variable and independent variables, the regression models are categorized into 4 types.

Types of Regression

Linear Regression

→ Simple Linear Regression : Predicts a continuous outcome variable based on a single independent variable.

→ Multiple Linear Regression : Predicts a continuous outcome variable based on multiple independent variables.

Nonlinear Regression

→ Polynomial Regression : models the connection b/w dependent variable and independent variable as an n^{th} degree polynomial.

→ Logistic Regression : Predicts the possibility of a categorical outcome (eg: yes/no, pass/fail)

Regularization Techniques

→ Ridge Regression : Addresses multicolinearity by adding a penalty term to the regression equation.

→ Lasso Regression : Performs feature selection by shrinking some coefficients to zero.

→ ElasticNet Regression : Combines the properties of Ridge and Lasso Regression.

Other Types :

→ Stepwise Regression : selects independent variables for the model in a stepwise manner.

→ Time Series Regression : Analyzes data points collected over time.

→ Poisson Regression : Models count data (ex: no. of occurrences)

Note: Use Regression analysis for one of two purposes : predict the value of the dependent variable when you know the independent variables or predict the effect of an independent variable on the dependent variable.

Challenges in Using Regression

While the types of regression analysis are powerful, they aren't foolproof. It includes

1) violating the assumptions of regression analysis
2) Overfitting with complex models.
3) Interpreting coefficients incorrectly.
4) Handling categorical data in linear models.

1) Simple Linear Regression

Assume that there is only one independent variable 'x'. If the relationship blw x (

independent variable) and y (dependent or output variable) is modeled by the relation,

$$y = a + bx$$

2) Multiple Regression

Assume that there are multiple independent variables say $x_1, x_2, x_3 \cdots x_n$ of the relationship blw independent variables 'x' and dependent or output variable 'y' is modeled by the relation,

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

3) Polynomial Regression

Assume that there is only one independent variable x. Of the relationship blw 'independent variables 'x' and dependent or o/p variable 'y' is modeled by the relation,

$$y = a_0 + a_1 * x + a_2 * x^2 + \cdots + a_n * x^n$$

for some +ve integer $n > 1$, then we have a polynomial regression.

$\Rightarrow n = 1 \rightarrow$ normal regression.

Classification in Machine Learning

Classification is defined as the process of predicting class or category from observed values or given data points. The categorized o/p can have the form such as Black or white or spam or no spam. It is a supervised learning technique where an algorithm is trained with labeled data to predict the labeled data to predict the category of new data.

Types of Learners:

We have two types of learners in respective to classification problems.

1) Lazy learners: These learners waits for the testing data to be appeared after storing the training data. classification is done only after getting the testing data. They spend less time on training but more time on predicting

Ex: KNN

2) Eager Learners: These learners construct classification model without waiting for the testing data to be appeared after storing the training data. They spend more time on training and less time on testing.

Ex: Decision trees, Naive Bayes and ANN.

Naive Bayes classification

It is a classification algorithm based on Bayes theorem. The algorithm assumes that the features are independent of each other, which is why it is called "naive". It calculates the probability of a sample belonging to a particular class based on probabilities of its features. Ex, a phone may be considered as smart if it has touch screen, internet facility, good camera etc. Even if all these features are dependent on each other, but all these features independently contribute to the probability of that the phone is a smart phone.

These classifiers are powerful and widely used in various machine learning tasks, particularly in text classification and sentiment analysis. It calculates the probability of a data point belonging to each class based on observed features. It then predicts the class with highest probability as final prediction.

Naive Bayes

→ This refers to Bayes' theorem, a theorem in probability theory that helps to calculate the conditional probability of an event (the class label) happening, given another event (the features) has already occurred.

→ They treat all features as independent of each other, given the class label. This means the presence or absence of one feature doesn't influence the presence or absence of another feature, considering the class is already known. In reality, this assumption isn't always true.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

where

$P(A|B)$ → Conditional probability of event A occurring given that the event B has occurred.

$P(B|A)$ → the probability of event B occurring given that the event A has already occurred.

$P(A)$ → prior ~~prolorat~~ probability of event A, before considering any new evidence.

$P(B)$ → probability of event B.

Person	COvid (Yes/No)	Flu (Yes/No)	Fever (Yes/NO
1	Yes	NO	Yes.
2	NO	Yes	Yes
3	Yes	Yes	Yes
4	No	NO	NO
5	Yes	NO	yes
6	NO	No	Yes
7	Yes	NO	Yes
8	Yes	NO	NO
9	NO	Yes	Yes
10	NO	Yes	NO

Given Person (flu, Covid)

$P(\text{yes} \mid \text{flu}, \text{covid}) = P(\text{flu}/\text{yes}) \times P(\text{Covid}/\text{Yes}) \times P(\text{Yes})$

$P(\text{No} \mid \text{flu}, \text{Covid}) = P(\text{flu}/\text{No}) \times P(\text{covid}/\text{NO}) \times P(\text{No})$

Step1 : Prior probability

$$P(\text{fever} = \text{yes}) = \frac{7}{10}$$ — yes in dataset / total values in dataset

$$P(\text{fever} = \text{no}) = \frac{3}{10}$$

Step2 : Conditional Probability :

	fever: yes	NO
Covid	3/7	2/3
Flu	4/7	2/3

$$P(\text{yes} \mid \text{flu}, \text{covid}) = P(\text{flu} \mid \text{yes}) * P(\text{covid} \mid \text{yes}) * P(\text{yes})$$

(yes: fever)

$$= \frac{3}{7} \times \frac{4}{7} \times \frac{7}{10} = \frac{126}{7035} = 0.17$$

$$P(\text{NO} \mid \text{flu}, \text{covid}) = P(\text{flu} \mid \text{NO}) * P(\text{covid} \mid \text{NO}) * P(\text{NO})$$

$$= \frac{2}{3} \times \frac{2}{3} \times \frac{3}{10} = \frac{4}{30} = \frac{2}{15} = 0.13$$

The person suffering from covid will suffer from fever.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	NO
D2	Sunny	Hot	High	Strong	NO
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	NO
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	NO
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	NO

New instance

(Outlook = Sunny, Temp = Cool, Humidity = high, Wind = Strong) – Classify whether a person enjoys tennis or not.

Prior Probability:

$P(\text{PlayTennis} = \text{yes}) = \frac{9}{14} = 0.64$

$P(\text{PlayTennis} = \text{no}) = \frac{5}{14} = 0.36$

} based on dataset

Current / conditional Probability:

Outlook	Y	N
Sunny	2/9	3/5
Overcast	4/9	0/5
rain	3/9	2/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5

Temp	Y	N
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Windy	Y	N
strong	3/9	3/5
weak	6/9	2/5

New instance (outlook = sunny, Temp = cool, Humidity = high, wind = strong)

$$V_{NB} = \underset{v_j \in \{yes, no\}}{\text{argmax}} \; P(v_j) \prod_i P(a_i | v_j)$$

(product)

$$= \underset{v_j \in \{yes, no\}}{\text{argmax}} \; P(v_j)$$

$$P(v_j) \Rightarrow P(\text{outlook} = \text{sunny} | v_j) \, P(\text{Temp} = \text{cool} | v_j) \, P(\text{Humidity} = \text{high} | v_j) \, P(\text{Wind} = \text{strong} | v_j)$$

$V_{NB}(yes) = P(yes)\, P(sunny|yes)\, P(cool|yes)$
$P(high|yes)\, P(strong|yes)$

$= \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{27}{7 \times 9 \times 9 \times 9}$

$= \frac{1}{7 \times 9 \times 3} = \frac{1}{189} = 0.00529$

$= 0.0053$

$V_{NB}(no) = P(no)\, P(sunny|No)\, P(cool|no)$
$P(high|no)\, P(strong|no)$

$= \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} =$

$= \frac{9 \times 4}{14 \times 25 \times 5} = \frac{36}{1750} = 0.02057$

$= 0.0206$

$V_{NB}(yes) = \frac{V_{NB}(yes)}{V_{NB}(yes) + V_{NB}(NO)} = \frac{0.0053}{0.0259}$

$= 0.2046 = 0.205$

$V_{NB}(no) = \frac{V_{NB}(no)}{V_{NB}(yes) + V_{NB}(NO)} = \frac{0.0206}{0.0259}$

$= 0.795$

The person will not playTennis as no is having highest probability.

Example - 2

1) Estimate conditional probabilities of each attributes {color, legs, height, smelly} for the species class: {M, H} using the table.

2) Using these probabilities estimate the probability values for the new instance.

(color = green, legs = 2, height = tall & smelly = NO)

NO	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	NO	M
3	Green	3	short	yes	M
4	White	3	short	Yes	M
5	Green	2	short	NO	H
6	White	2	Tall	NO	H
7	White	2	Tall	NO	H
8	White	2	short	c/es	H

Prior Probability

$$P(M) = \frac{4}{8} = 0.5 \qquad P(H) = \frac{4}{8} = 0.5$$

Conditional Probability

Color	M	H
White	2/4	3/4
Green	2/4	1/4

Legs	M	H
2	1/4	4/4
3	3/4	0/4

Height	M	H
Tall	3/4	2/4
Short	1/4	2/4

Smelly	M	H
yes	3/4	1/4
NO	1/4	3/4

New instance (color = Green, legs = 2, height = tall, smelly = NO)

$$V_{NB} = \underset{v_j (M,H)}{\operatorname{argmax}} \; v_j \; \Pi_j \; P(a_i | v_j)$$

$$V_{NB}(M) = P(M)\, P(\text{Green}|M)\, P(2|M)\, P(\text{tall}|M)\, P(\text{NO}|M)$$

$$= \frac{4}{8} \times \frac{2}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{3}{16 \times 16} = \frac{3}{256}$$

$$= 0.01171$$

$$V_{NB}(H) = P(H)\, P(\text{Green}|H)\, P(2|H)\, P(\text{tall}|H)\, P(\text{No}|H)$$

$$= \frac{4}{8} \times \frac{1}{4} \times \frac{4}{4} \times \frac{2}{4_2} \times \frac{3}{4} \qquad = \frac{3}{8 \times 4 \times 2}$$

$$= \frac{3}{64} \qquad = 0.047$$

$$V_{NB}(H) > V_{NB}(M)$$

Hence new instance belongs to species H

Example - 3

Example No	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

New Instance :- (Red, SUV, Domestic) → (Yes/No)

Prior Probability

$$P(\text{Yes}) = \frac{5}{10} = 0.5$$

$$P(\text{No}) = \frac{5}{10} = 0.5$$

Conditional Probability

Color	Yes	NO
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	NO
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	NO
Domestic	2/5	3/5
Imported	3/5	2/5

$$V_{NB}(\text{yes}) = P(\text{yes})\ P(\text{Red}|\text{yes})\ P(\text{SUV}|\text{Yes})\ P(\text{Domestic}|\text{Yes})$$

$$= 0.5 \times \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} = \frac{0.5 \times 6}{125} = \frac{3}{125}$$

$$= 0.024$$

$$V_{NB}(\text{no}) = P(\text{no})\ P(\text{Red}|\text{no})\ P(\text{SUV}|\text{no})\ P(\text{Domestic}|\text{no})$$

$$= 0.5 \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{3 \times 3}{25 \times 5} = \frac{9}{125}$$

$$= 0.072$$

Hence new instance belongs to "No".

Example - 4

Consider a football game between two rival teams, say team A and team B. Suppose team A wins 65% of the time and team B wins the remaining matches. Among the games won by team A, only 35% of them comes from playing at team B's football field. On the other hand, 75% of victories for team B are obtained while playing at home.

1) If team B is to host the next match blw the two teams, what is the probability that it will emerge as the winner?

2) If team B is to host the next match blw the two teams, who will emerge as the winner?

Y - winning football match

X - Hosting " "

Probability that team A wins is

$P(Y_A) = 0.65$

Probability that team B wins is

$P(Y_B) = 1 - P(Y_A) = 0.35$

Probability that team A hosted match it had won is $P(X_A | Y_B) = 0.75$

probability that team B hosted the match won by team A is $P(X_B | Y_A) = 0.35$

1. $P(Y_B | X_B) = \dfrac{P(X_A | Y_B) \times P(Y_B)}{P(X_B)}$

$= \dfrac{P(X_A | Y_B) \times P(Y_B)}{P(X_B | Y_A) P(Y_A) + P(X_A | Y_B) P(Y_B)}$

$= \dfrac{0.75 \times 0.35}{(0.35 \times 0.65 + 0.75 \times 0.35)}$

$= 0.5357$

(1)

$$P(Y_A \mid X_B) = \frac{P(X_B \mid Y_A)\, P(Y_A)}{P(X_B)}$$

$$= \frac{P(X_B \mid Y_A) \times P(Y_A)}{P(X_B \mid Y_A)\, P(Y_A) + P(X_B \mid Y_B)\, P(Y_B)}$$

$$= \frac{0.35 \times 0.65}{0.35 \times 0.65 + 0.75 \times 0.35} = 0.4642$$

∴ B will emerge as winner

Example - 5

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

1) Estimate the conditional probabilities of $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$ $P(C|-)$

2) Use the conditional probability estimates and predict the class label for the test sample $P(A=0, B=1, C=0)$

Gaussian Naive Bayes

It is a type of Naive Bayes method working on continuous attributes and the data features that follows Gaussian distribution throughout the dataset. Gaussian distribution is also called normal distribution. It is a statistical model that describes the distributions of continuous random variables in nature and is defined by its bell-shaped ~~Cave~~ curve. The two important features are mean (μ) and standard deviation (σ). The mean is the average value of a distribution, and the standard deviation is the "width" of the distribution around the mean.

It assumes that the likelihood ($P(x_i|y)$) follows Gaussian Distribution for each x_i within y_k.

$$\therefore P(x_i|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

x_i – feature value

μ – mean of feature values for class y_k

σ – Standard deviation

π – const

e – base

It is effective for continuous data because it assumes each feature follows a Gaussian (normal) distribution. When this assumptions holds true the algorithm performs well.

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

Based on the following data determine the gender of a person having height 6 ft, weight 130 lbs and foot size 8 inch use Naive Bayes Algorithm.

$$P(\text{Male}) = \frac{4}{8} = 0.5$$

$$P(\text{Female}) = \frac{4}{8} = 0.5$$

Male:

$$\text{Mean (Height)} = \frac{(6 + 5.92 + 5.58 + 5.92)}{4} = 5.855$$

$$\text{Variance (Height)} = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

$$= \frac{(6 - 5.855)^2 + (5.92 - 5.855)^2 + \cdots (5.58 - 5.855)^2 + (5.92 - 5.855)^2}{4-1}$$

$$= 0.035055$$

Mean (weight)

Variance (Waght)

Mean (Foot size)

Variance (Foot size).

Hardly Calculate for Female also

Sex	Mean (height	Variance (height)	Mean (weight)	Variance (weight	Mean (foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New instance to be classified is

Sex	Height	weight	Foot size
Sample	6	130	8

$$\text{Posterior (Male)} = \frac{P(M) \times P(H/M) \times P(W/M) \times P(FS/M)}{\text{Evidence}}$$

(Male, Height, weight, Foot size)

$$\text{Posterior (Female)} = \frac{P(F) \times P(H/F) \times P(W/F) \times P(FS/F)}{\text{Evidence}}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

σ^2 – Variance

μ – Mean.

$$P(H|M) = \frac{1}{\sqrt{2 \times 3.142 \times 0.035033}} \times e^{-\frac{(6-5.855)^2}{2 \times 0.035033}}$$

$$= 1.5789$$

$$= \frac{1}{0.155667} \times e^{-\frac{(0.021025)}{0.070066}}$$

$$= \frac{1}{0.0878} \times e^{-\left(\frac{0.021025}{0.07066}\right)^2}$$

$$= 11.39 \times e^{-0.300} =$$

$$11.39 \times 0.7408 = 1.5789$$

$$P(W|M) = 5.9881e^{-6}$$

$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$

$$P(W|F) = 1.6789e^{-2}$$

$$P(FS|F) = 2.8669e^{-1}$$

Posterior (Male) = $0.5 \times 1.5789 \times 5.9881e^{-6} \times 1.3112e^{-3}$

$= 6.1984e^{-9}$

Posterior (Female) = $0.5 \times 2.2346e^{-1} \times 1.6789e^{-2} \times 2.8669e^{-1} = 5.377e^{-4}$

Female probability is more when compared to male. So new instance is classified as Female.

CART (Classification and Regression Tree) is a predictive algorithm in Machine learning that models how the target variable can be predicted based on i/p features

Decision trees in the following categories are together referred to as CART;

→ Classification trees: Used when the target variable is categorical. These trees predict the class or category to which the i/p data most likely belongs.

→ Regression trees: Used when the target variable is continuous. These trees predict

a numerical value based on I/P features

Working:

1) Tree Structure:

CART constructs a binary tree structure where each internal node represents a decision based on a feature's value, and each leaf node corresponds to a class label (for classification) or a predicted value (for regression)

2) Splitting Criteria:

At each node, CART selects the feature and the value to split the dataset into two subsets. It chooses the split that maximizes the homogeneity or purity of the resulting subsets, measured by metrics like Gini impurity for classification or mean squared error for regression.

Calculating Gini index

The Gini index is a metric for classification tasks in CART. It stores the sum of squared probabilities of each class.

The probability that a particular variable would be misclassified when chosen at

random is determined by Gini index, also referred to as Gini coefficient or Gini impurity.

Gini = 0 : Node is completely pure (only one class)

Gini = 1 : Node contains a mix of many classes and is highly impure.

Gini = 0.5, the elements are distributed evenly among many classes

How does CART for Classification Work:

→ CART for classification works by recursively splitting the training data into smaller and smaller subsets based on certain criteria.

→ The goal is to split the data in a way that minimizes the impurity within each subset.

→ Impurity is a measure of how mixed up the data is in a particular subset. For classification tasks, CART uses Gini impurity.

→ It measures the probability of misclassifying a random instance from a subset labeled according to the majority class. Lower Gini impurity means more purity of the subset.

The CART algorithm evaluates all potential splits at every node and chooses the one that best decreases the Gini impurity of the resultant subsets

This process continues until a stopping criterion is reached, like a maximum tree depth or a minimum number of instances in a leaf node.

How does CART works for regression

Regression CART works by splitting the training data recursively into smaller subsets based on specific criteria. The objective is to split the data in a way that minimizes the residual reduction in each subset.

Residual reduction is a measure of how much the average squared difference blw the predicted values and the actual values for the target variable is reduced by splitting the subset.

The lower the residual reduction, the better the model fits the data. CART evaluates every possible split at each node and selects the one that results in the greatest reduction of residual error in the resulting subsets.

This process is repeated until a stopping criteria is met, such as reaching the maximum tree depth or having too few instances in a leaf node.

Day	Outlook	Temp	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	0
2	Sunny ~~Overcast~~	Hot	High	Strong	0
3	~~Rain~~ Overcast	Hot	High	Weak	1
4	Rain	Mild	High	Weak	1
5	Rain	Cool	Normal	Weak	1
6	Rain	Cool	Normal	Strong	0
7	Overcast	Cool	Normal	Strong	1
8	Sunny	Mild	High	Weak	0
9	Sunny	Cool	Normal	Weak	1
10	Rain	Mild	Normal	Weak	1
11	Sunny	Mild	Normal	Strong	1
12	Overcast	Mild	High	Strong	1
13	Overcast	Hot	Normal	Weak	1
[illegible]	Rain	Mild	High	Strong	0

CART using regression

Mean = $\frac{\text{sum}}{\text{Total}}$

Variance : $\frac{1}{n} \Sigma (y_i - \bar{y})^2$

~~Outlook~~

Target variable : Decision (0 or 1)

Total records = 14

Mean = $\frac{\Sigma \text{Decision}}{14}$

Count of 1's = 9

" " 0's = 5

Mean = $\frac{9}{14} = 0.643$

Compute MSE

$$MSE = \frac{1}{n} \Sigma (y_i - \bar{y})^2$$

$$= \frac{1}{14} [9(1 - 0.643)^2 + 5(0 - 0.643)^2]$$

$$= \frac{1}{14} [9(0.127) + 5(0.413)]$$

$$= \frac{1}{14} [1.143 + 2.065]$$

$$= 0.229$$

Split each attribute

1 Split outlook

Sunny - 5, Overcast - 4, Rain - 5

a) outlook = Sunny

Decisions : 0, 0, 0, 1, 1

Mean $\bar{y} = \frac{2}{5} = 0.4$

MSE $= \frac{1}{5}\left[3(0-0.4)^2 + 2(1-0.4)^2\right] = 0.24$

b) Outlook = Overcast

Decisions : 1, 1, 1, 1

Mean = 1

MSE = 0

c) Outlook = Rain

Decisions : 1, 1, 0, 1, 0

Mean = $\bar{y} = \frac{3}{5} = 0.6$

MSE $= \frac{1}{5}\left[3(1-0.6)^2 + 2(0-0.6)^2\right] = 0.24$

Weighted MSE for outlook

$$MSE = \frac{5}{14}(0.24) + \frac{4}{14}(0) + \frac{5}{14}(0.24)$$

$$= 0.171$$

Outlook = 0.171

11 Split of Temperature

Hot - 4 Cool - 4 Mild - 6

a) Temp = Hot

0, 0, 1, 1

Mean : $\bar{y} = \frac{1}{2} = 0.5$

$MSE = \frac{1}{4}\left[2(0-0.5)^2 + 2(1-0.5)^2\right]$

$= \frac{1}{4}\left[2(0) + 2(0.5)^2\right]$

$= \frac{1}{4}\left[2(0.25)\right] = \frac{1}{4}(0.50)$

0.5
0.0
0.5

1.0
0.5

$MSE = \frac{1}{4}\left[(0-0.5)^2 + (0-0.5)^2 + (1-0.5)^2 + (1-0.5)^2\right] = 0.25$

b) Temp = Mild

Decisions : 1, 0, 1, 1, 1, 0

Mean = $\bar{y} = \frac{4}{6} = 0.667$

$MSE = \frac{1}{6}\left[2(0-0.667)^2 + 4(1-0.667)^2\right]$

$= 0.222$

c) Temp = Cool

1, 0, 1, 1

Mean $\bar{y} = \frac{3}{4} = 0.75$

$$MSE = \frac{1}{4}\left[1(0-0.75)^2 + 3(1-0.75)^2\right] = 0.188$$

$$\text{Weighted MSE} = \frac{4}{14}(0.25) + \frac{6}{14}(0.222) + \frac{4}{14}(0.188)$$

$$= 0.071 + 0.095 + 0.054 = 0.220$$

Temp = 0.220

III Split of Humidity

Decisions: 0, 0, 1, 1, 0, 1, 0

Humidity = High

$$\text{Mean} = \frac{3}{7} = 0.43$$

$$MSE = 0.245$$

Humidity = Normal

$$\text{Mean} = \frac{6}{7} = 0.86$$

$$MSE = 0.122$$

$$\text{Weighted MSE} = \frac{7}{14}(0.245) + \frac{7}{14}(0.122) = 0.184$$

Humidity = 0.184

IV Split of Wind

Decision – Weak – 1, 1, 1

Mean = 1

MSE = 0

Wind = Strong

Decision : 0, 0

Mean = 0 MSE = 0

IV Split of Wind

Decision - Wind - weak - 0, 1, 1, 1, 0, 1, 1, 1

Mean : $\frac{6}{8} = 0.75$

MSE : $\frac{1}{8}[6(1-0.75)^2 + 2(0-0.75)^2$

$= \frac{1}{8}[6(0.0625) + 2(0.5625)] = 0.1875$

Decision - Wind - Strong - 0, 0, 1, 1, 1, 0

Mean : $\frac{3}{6} = 0.5$

MSE : $\frac{1}{6}[3(1-0.5)^2 + 3(0-0.5)^2$

$= \frac{1}{6}[3(0.25) + 3(0.25)]$

$\frac{1}{6}[2(0.75)] = \frac{1.5}{6} = 0.25$

Tot weighted MSE =

$\frac{8}{14}(0.1875) + \frac{6}{14}(0.25)$

$= 0.1071 + 0.1071 = 0.2142$

Wind : 0.2142

Outlook	= 0.171
Temp	= 0.220
Humidity	= 0.184
Wind	= 0.2142

Outlook is best.

Outlook

Overcast = 0 — leaf node.

Outlook — Sunny

Day	Outlook	Temp	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	0
2	Sunny	Hot	High	Strong	0
8	Sunny	Mild	High	Weak	0
9	Sunny	Cool	Normal	Weak	1
11	Sunny	Mild	Normal	Strong	1

Split: Temp : Hot

Decision — 0, 0

Mean = 0

MSE = 0

Cool : Decision — 1

Mean = 1

$MSE = \frac{1}{14}(1-1) = 0$

Mild — Decision — 0, 1

$Mean = \frac{1}{2} = 0.5$

$MSE = \frac{1}{2}(1(1-0.5)^2 + 0(0-0.5)^2)$

$= \frac{1}{2}(1(0.25))$

$= 0.125$

Tot MSE : $\frac{1}{14}(0.125) = 0.00892$

Humidity

High

Decision : 0, 0, 0

Mean : 0 , MSE : 0

Normal

Decision : 1, 1

Mean : $\frac{2}{2} = 1$

MSE : $\frac{1}{2}[2(1-1^2] = 0$

Tot MSE = 0

Wind

Weak

Decision : 0, 0, 1

Mean : $\frac{1}{3} = 0.3$

MSE $= \frac{1}{3}[2(0-0.3)^2 + 1(1-0.3)^2]$

$= \frac{1}{3}[2(0.09) + 1(0.49)] = 0.223$

Strong

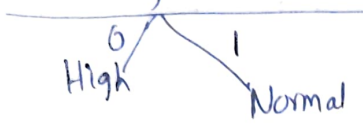
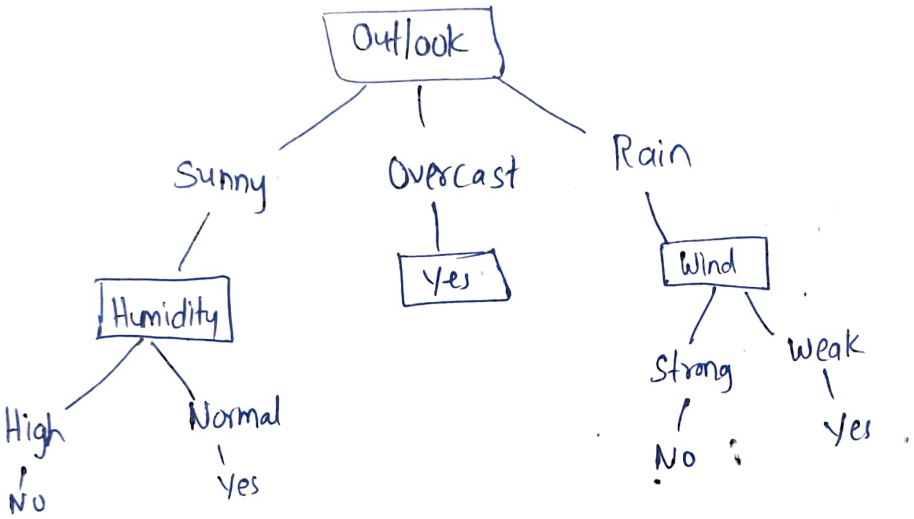
Decision : 0, 1

Mean : $\frac{1}{2} = 0.5$

$$MSE = \frac{1}{2}\left[1(1-0.5)^2 + 1(0-0.5)^2\right]$$

$$= \frac{1}{2}\left[0.25 + 0.25\right] = 0.25$$

Humidity is taken



Nearly do ~~with~~ for all attributes.